

University of Groningen

High-Dimensional Astronomical Data Using Dimension Reduction

Kim, Youngjoo; Telea, Alexandru; Trager, Scott

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Kim, Y., Telea, A., & Trager, S. (2018). *High-Dimensional Astronomical Data Using Dimension Reduction*. Poster session presented at XXX Canary Islands Winter School of Astrophysics, Tenerife, Spain.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

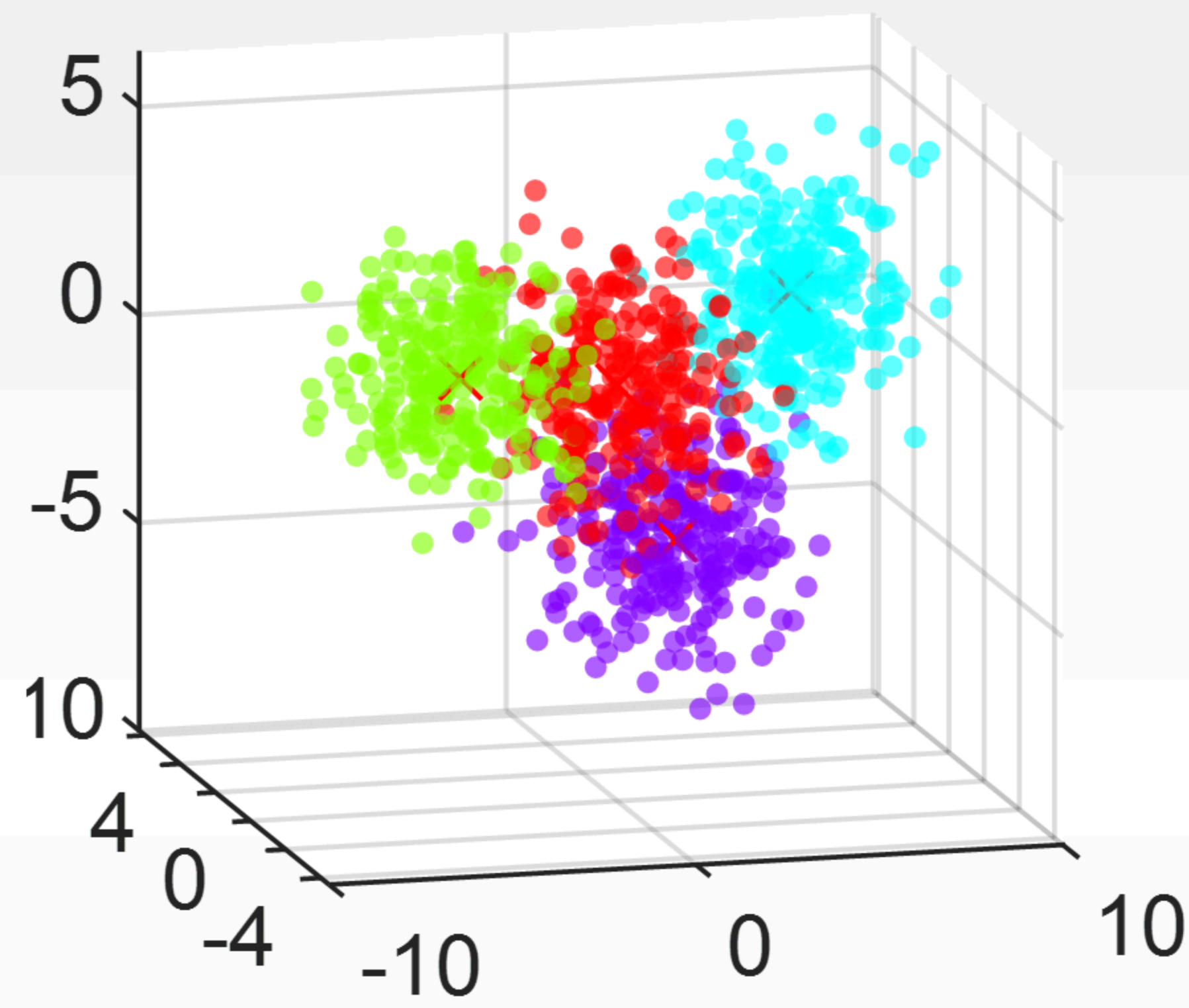
Big Data Analysis in Astronomy

Visual Analytics for High-Dimensional Astronomical Data Using Gradient Clustering in Dimensionality Reduction

Youngjoo Kim¹ Alexandru C. Telea¹ Scott C. Trager¹

¹Faculty of Science and Engineering, University of Groningen

Original Data



Gaussian random data with
four clusters in N-D

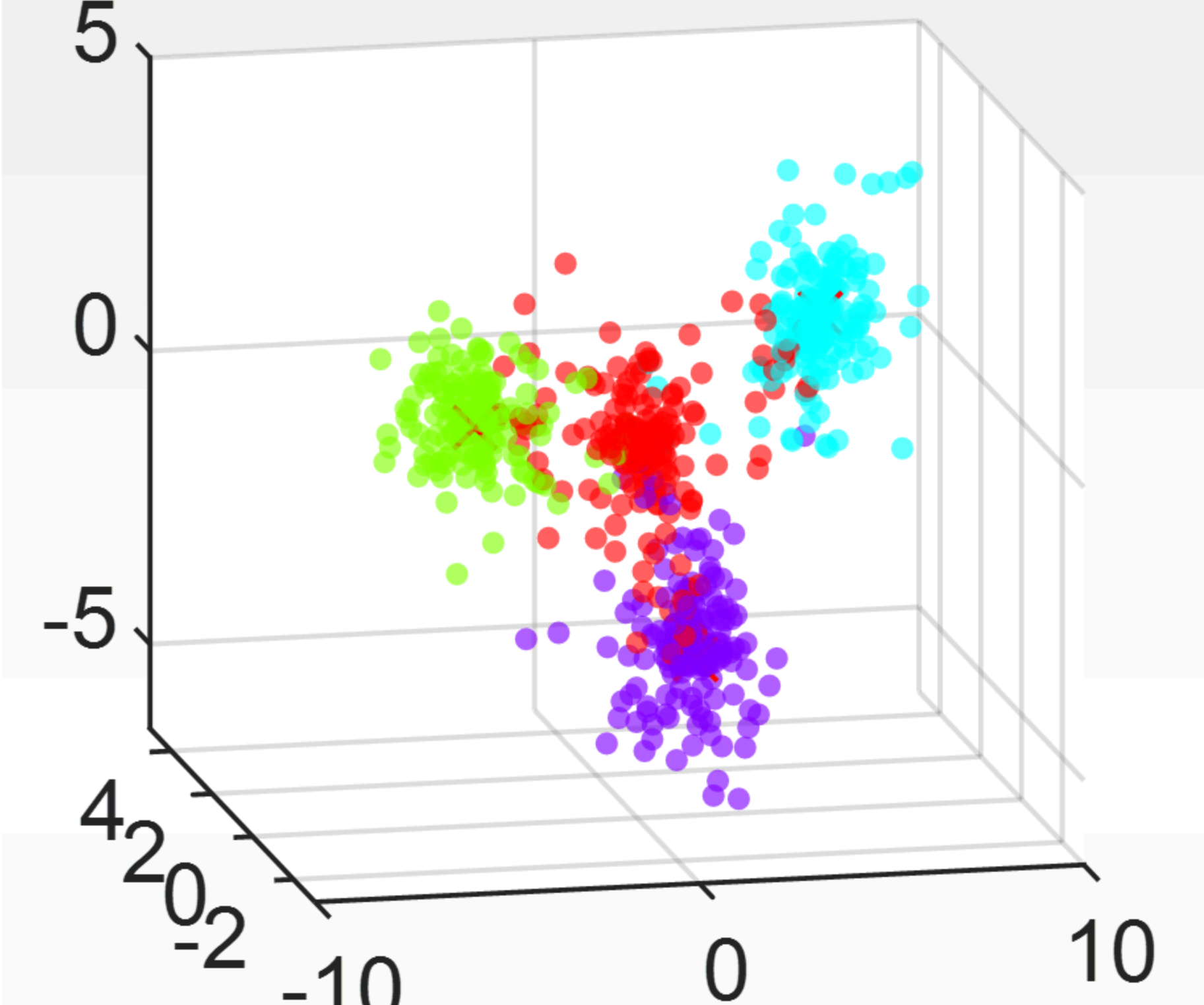
Preprocess high-dimensional data

“Whoop”

Local Gradient Clustering (LGC)

Shift points along the gradient of the kernel
density estimator

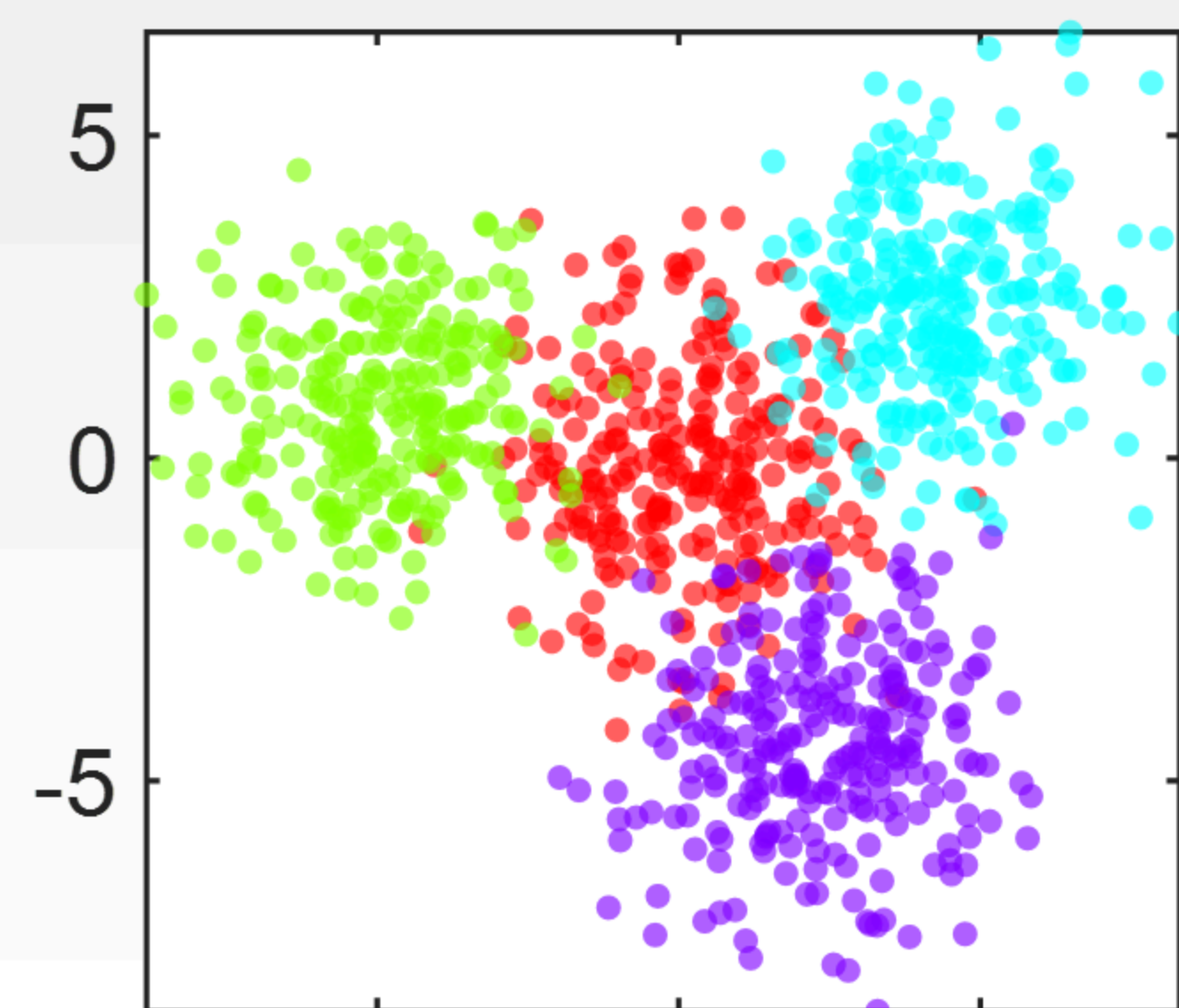
Preprocessed Data



Clusters are separated in N-D

Dimensionality Reduction

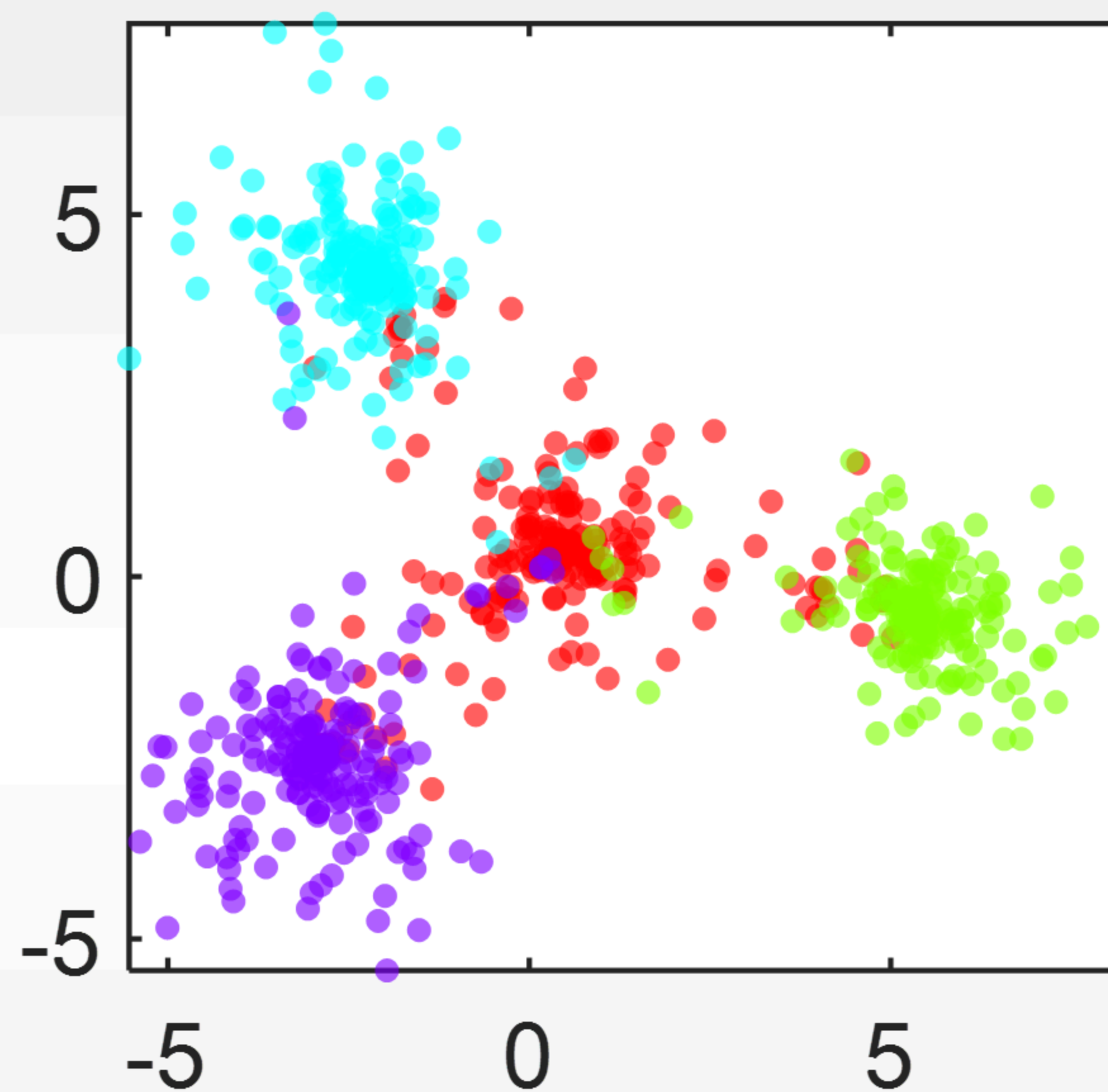
LGC+LMDS



Clusters are not well separated
in the 2D projection

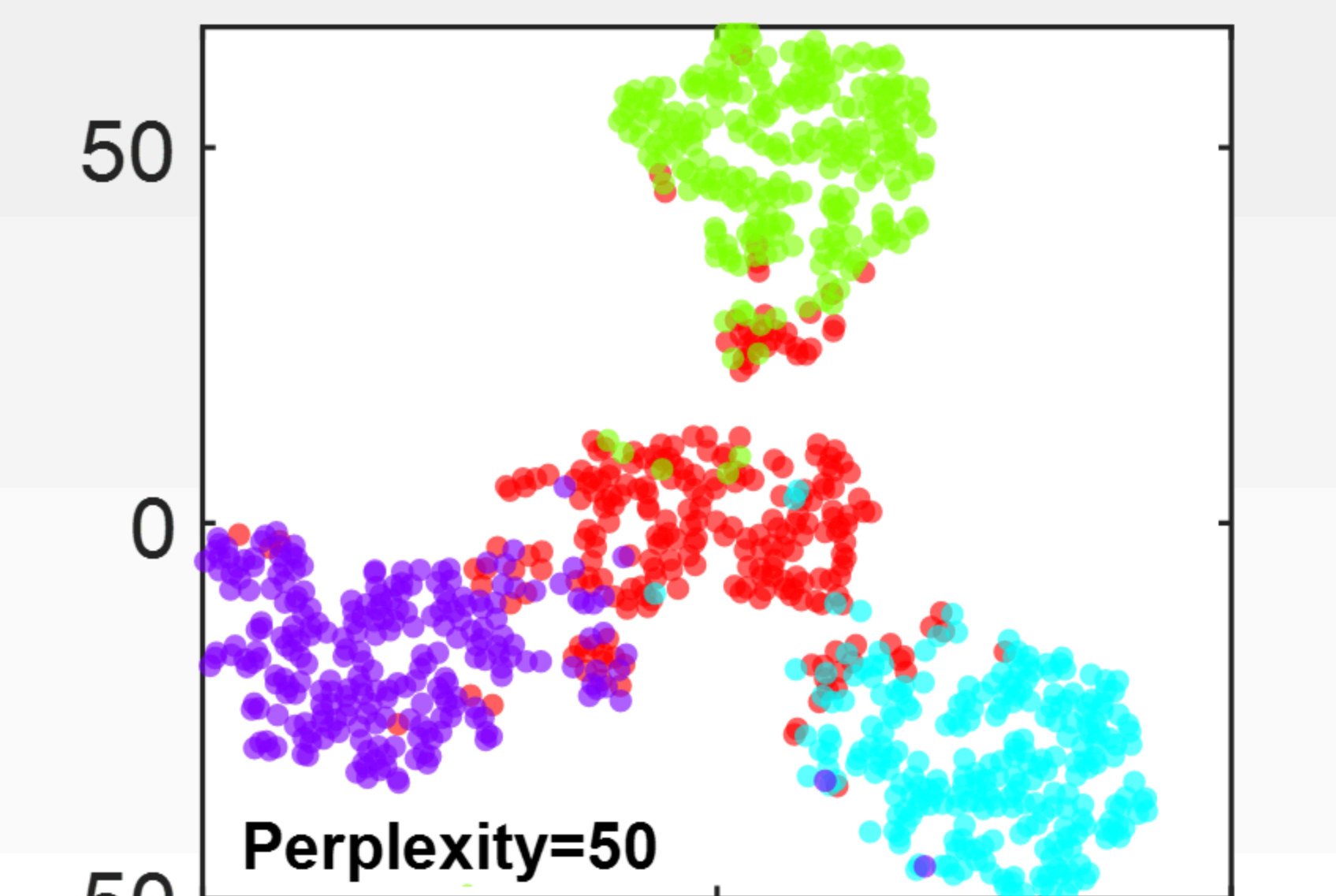
Landmark Multidimensional Scaling
(LMDS) [1]

VS.



Clusters are well separated in the 2D projection

VS.



Clusters are well separated
in the 2D projection

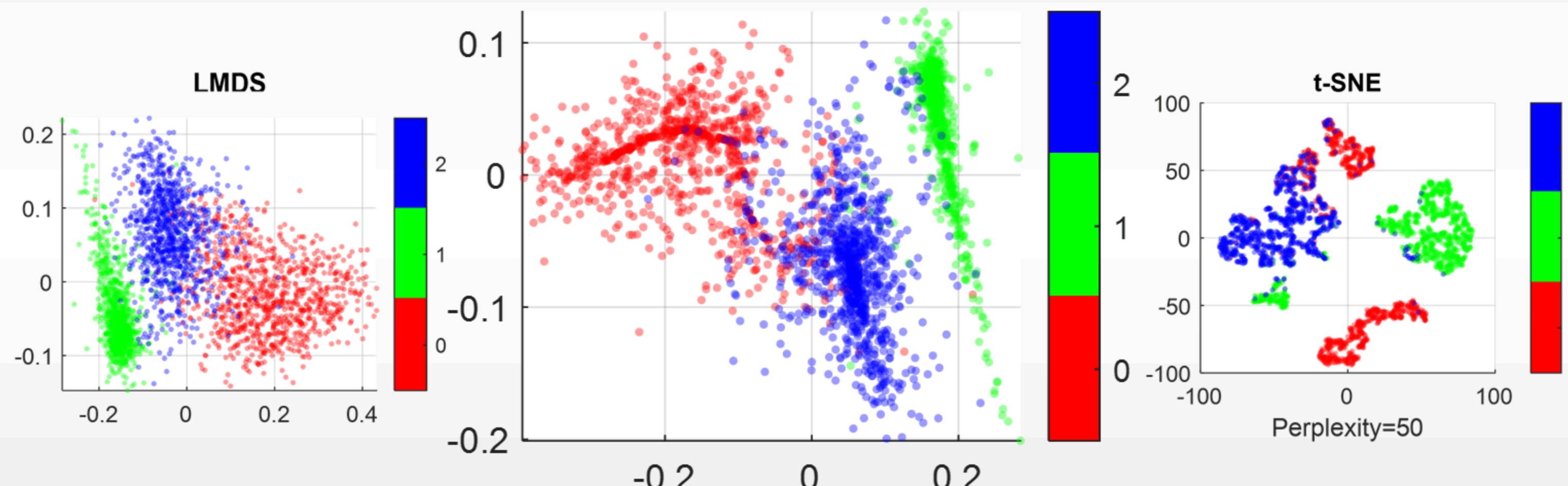
t-Stochastic Neighbor Embedding
(t-SNE) [2]

Apply to MNIST Dataset

- First three digits (0, 1, and 2) of the MNIST dataset [3]
 - Two Vertical/horizontal gradient histogram features
 - Three Mean-centered pixel coordinate features

0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2

LMDS with LGC



Summary

Preprocess high-dimensional data so that potential clusters are well separated after dimensionality reduction

Method

- Generate a kernel density estimator with an Epanechnikov kernel and use only the neighbors when computing the density [4-5]
- Shift points along the gradient of the kernel density estimator as in gradient ascent → points moving towards the center → contraction [6]
- Perform LMDS [1]

Cluster Separation & Scalability & Predictability

- Clusters are well separated after the projection by preprocessing the data with local-based gradient clustering
- More computationally scalable than t-SNE, in terms of wall-clock time
- Predictable outcome with three parameters

Future Work

- Apply to astronomical surveys such as GAIA DR2 and GALAH DR2 to find underlying patterns in the data

References

- [1] V. De Silva and J. B. Tenenbaum, "Sparse multidimensional scaling using landmark points," Technical report, Stanford University, Vol. 120, 2004.
- [2] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of machine learning research, No. 9, pp. 2579-2605, 2008.
- [3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.
- [4] M. Muja and D. G. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration," International Conference on Computer Vision Theory and Applications (VISAPP'09), 2009.
- [5] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," Theory of Probability and its Applications, Vol. 14, No.1, pp. 153-158, 1969.
- [6] K. Fukunaga and L. Hostetter, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Transactions on information theory, Vol. 21, No. 1, pp. 32-40, 1975.